DOCUMENT RESUME

ED 307 312 TM 013 402

AUTHOR De Ayala, R. J.; And Others
TITLE A Comparison of the Graded Response and Partial Credit Models for Assessing Writing Ability.
PUB DATE Mar 89
NOTE 26p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, March 28-30, 1989).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; Essay Tests; *Holistic Evaluation; Interrater Reliability; Latent Trait Theory; Models; *Scoring; Secondary Education; *Secondary School Students; Writing (Composition); *Writing Evaluation
IDENTIFIERS *Graded Response Model; *Partial Credit Model; Writing Samples

ABSTRACT

The graded response (GR) model of Samejima (1969) and the partial credit model (PC) of Masters (1982) were fitted to identical writing samples that were holistically scored. The performance and relative benefits of each model were then evaluated. Writing samples were both expository and narrative. Data were from statewide assessments of secondary school students' writing ability for 1985 through 1988, for a total of 2,000 examinees. An examinee's four samples were randomly given to a team of 80 to 100 trained raters. Results indicate that both models were useful for the calibration of writing samples. For this item set, the GR model provided more information than did the PC model for both the rating scales examined. In some cases, one might prefer the PC model because of the fewer parameters to estimate and the minimal gains to be expected by using the GR model in this context. It is possible, if data collection is structured appropriately, to perform an interrater agreement analysis through the use of item or test information functions. The advantages of item response theory methods may be realized with essay-type examinations. Eleven graphs are provided. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED307312

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

RALPH DE AYALA

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A COMPARISON OF THE GRADED RESPONSE AND PARTIAL CREDIT MODELS FOR ASSESSING WRITING ABILITY

R.J. De Ayala, University of Maryland
B.G. Dodd & W.R. Koch, University of Texas

TM013402

BEST COPY AVAILABLE

A Comparison of the Graded Response and Partial Credit Models for Assessing Writing Ability[1]

R.J. De Ayala, University of Maryland
B.G. Dodd & W.R. Koch, University of Texas

The assessment of writing proficiency may be accomplished either through direct methods (i.e., the actual demonstration of the skill) or by indirect methods (e.g., through the use of objective exams). Interest in the direct measurement of writing proficiency has been growing and half of all statewide writing assessment programs now rely exclusively on writing samples; the remaining programs use both direct and indirect assessments (Stiggins & Bridgeford, 1983). This study was concerned with the direct method of assessing writing proficiency.

The direct assessment of writing proficiency may be performed either through the holistic approach or the analytic method. In the analytic scoring method the ideal answer is decomposed into a set of components each of which is assigned a specific number of points. Assessment requires that the rater evaluate the examinee's writing with respect to each component and assign points according to the perceived quality of the writing on a given component. The examinee's score is a linear composite of his or her component scores.

In contrast, the holistic technique requires that the rater evaluate each writing sample with respect to an ideal answer or standard; multiple standards which are less than the ideal answer and which vary across the quality continuum may be used as additional standards. The examinee's score is typically a single score which is an assessment of the rater's global impression of the quality of the written piece with respect to the standard(s).

[1]Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, March, 1989.

TM013402

In order to improve the reliability of the writing assessment, multiple raters may be used to evaluate the same writing sample and their ratings pooled to form a single rating (e.g., the examinee's score is the average or the sum of the raters' ratings). A more complete discussion of the issues involved in the direct assessment of writing ability may be found in Breland (1983).

The above methods have historically been and currently are approached primarily through classical test theory. A few researchers have recently begun to approach writing assessment using item response theory (IRT) models (e.g., Pollitt & Hutchinson, 1987; Ackerman, 1986). However, although the results of these studies were encouraging, the comparative advantages and disadvantages of the IRT models used could not be assessed due to methodological differences. For instance, although both studies assessed writing proficiency by the analytical method, Ackerman's method used five components (e.g., paragraph development, spelling), whereas the Pollitt and Hutchinson study used a different set of three components (e.g., appropriacy, ideas).

An additional difference between the studies was Ackerman's use of one expository question decomposed into 15 'items' (i.e., 3 raters X 5 components), whereas the Pollitt and Hutchinson study used five writing tasks rated on three components to produce 15 'items'; both studies used secondary school students and their teachers as the raters. In each study the 'items' were considered to be locally independent.

This study fitted Samejima's (1969) graded response (GR) model and Masters' (1982) partial credit (PC) model to identical writing samples which were holistically scored and then evaluated the performance and relative benefits of each model. Further, the writing samples used consisted of two classes of questions : expository and narrative. It was felt that the expository questions would be found to provide more information for higher ability examinees than the narrative items would because the expository items appeared to require greater discourse competence than did the narrative items; discourse

competence is defined as the selection and structuring of ideas with respect to the purpose of the writing task and the needs of the reader (Pollitt and Hutchinson, 1987). A third factor investigated was the influence of two different methods of pooling the raters' ratings on item parameter estimation.

Model Descriptions

The two polychotomous models, the GR and PC, are appropriate for items with ordered responses, such as attitude questionnaires and aptitude or achievement test items whose alternatives are inherently ordered or have been ordered according to degree of correctness (e.g., through partial credit scoring). In addition, ratings data (e.g., ratings of writing samples) may also be fitted by either model.

The GR model is a direct extension of the two-parameter model. As a result, the GR model contains a parameter which allows an assessment of an item's capacity to discriminate among examinees. In the GR model the examinee responses to item i are categorized into $m_i + 1$ categories, where higher categories indicate more of θ and m_i is the number of categories. Associated with each category of item i is a category score, x_i, with values $0..m_i$. The GR model may be expressed as :

$$P_{x_i}(\theta) = \frac{e^{Da_i(\theta - b_{x_i})}}{1 + e^{Da_i(\theta - b_{x_i})}} \qquad (1),$$

where θ is the latent trait, a_i is the discrimination parameter for item i, b_{x_i} is the difficulty parameter for category score x for item i, and the scaling constant D equals 1.702. P_{x_i} is the probability, p_{x_i}, of the examinee responding in category score x_i or higher for a given item; the probability of responding in the lowest category (i.e., $P_0(\theta)$) or higher is defined as 1.0. For instance, for an item with four response categories $P_2(\theta)$ is the probability of responding in categories 2 or 3 rather than in categories 0 or 1. Because P_{x_i} is the probability of responding in x_i or higher, the probability of responding in a particular category equals the difference between cumulative probabilities for adjacent categories (e.g., $p_2(\theta) = P_2(\theta) - P_3(\theta)$). When an item consists of

two categories (correct and incorrect), the GR model reduces to the two-parameter model.

In contrast to the GR model, the PC model provides a direct expression of the probability of an examinee with ability θ responding in a particular category. In the PC model the examinee-item interaction is modeled as :

$$P_{x_j}(\theta) = \frac{e^{\sum_{j=0}^{x_j}(\theta - b_{x_j})}}{\sum_{k=0}^{m_i} e^{\sum_{j=0}^{k}(\theta - b_{x_j})}} \qquad (2),$$

where θ is the latent trait, b_{x_j} is the difficulty parameter of the step associated with category score x_j of item i with m_i categories, where $x_j = 1..m_i$. A category score reflects the number of successfully completed steps. A "step" is simply a stage required to complete an item. For instance, the problem $((6/3)+2)^2$ is considered to contain three steps because there are three separate stages which must be completed (in a specific order) to correctly answer the problem (i.e., step 1 : 6/3, step 2 : the addition of 2 to the quotient, and step 3 : the squaring of the quantity). For notational convenience $\Sigma(\theta - b_{x_j})$ where $j=0$ is defined as being equal to zero.

Because the PC model is an extension of the Rasch model it assumes that all items are equally good at discriminating among examinees. In addition, as a member of the Rasch family, the PC model's item and person parameters may be estimated on the basis of the existence of sufficient statistics. Specifically, an examinee's test score contains all the information for estimating his or her ability and the items' difficulties may be estimated from a simple count of the number of persons completing each "step" of an item. Unlike the GR model, the PC model requires that the steps within an item be completed in sequence, although the steps need not be equally difficult nor be ordered in terms of

difficulty. If an item consists of only two categories, then the PC model reduces to the Rasch model.

Method

Data : The data came from a state-wide assessment of secondary school students' writing ability. This test has been given annually since its inception in 1984 and is required for graduation from high school. Except for the 1984 administration, the writing test consisted of four writing samples, two of which were expository in nature and the two remaining items were narrative; the 1984 administration contained one expository and one narrative item. Two of the four items used in the 1985 to 1988 administrations were from the 1984 testing. These two items appeared in all administrations and served as a 'link' across administrations so that the separate testings could be placed on the same scale.

An examinee's four writing samples were randomly given to a team of 80 to 100 specially trained raters. The ratio of writing samples to individual rater, made it very unlikely that the same rater would rate more than one writing sample by a given examinee.

Each writing sample was holistically scored by two raters on a 1 to 4 scale; items with 0 scores indicated that the writing sample could not be scored (e.g., the student did not provide an answer) and were eliminated from analysis. If the two raters' ratings disagreed by more than two points a third rater was used. Exact interrater agreement occurred on approximately 76% of the ratings and periodic "check packets" (i.e., pre-scored writing samples) were distributed to monitor any drift in the ratings.

Rating Type : Because each writing sample had at least two ratings the impact of using a simple sum rating versus an average rating was investigated. The former method consisted of the sum of the raters' ratings (in the case of three raters the two closest ratings were used) which was then transformed to a 0 to 6 range (a.k.a., the sum rating scale). That is, the sum of the two ratings (range 2 to 8) was transformed by simply recoding a rating of 2 to 0, a rating of 3 to 1, etc. The second approach was to round the average of the two ratings to the nearest integer. In this latter method the original readers' 1 to 4 ratings

were recoded to 0 to 3 (i.e., a rating of 1 was recoded to 0, a rating of 2 was recoded to 1, etc.) before calculating the average and rounding to the nearest integer; this method will be known as the average rating scale.

Calibration : The MULTILOG 5.1 (Thissen, 1988) calibration program was used to fit both the GR and the PC models to the ten item pool. The use of a single calibration program for both models controlled for differences in the implementation of estimation algorithms when different calibration programs are used. Although MULTILOG provides for direct specification of the GR model, obtaining item parameter estimates for the PC model is not direct. Estimates for the PC model were obtained by imposing triangular contrasts on Bock's (1972) nominal response (NR) model (cf., Thissen & Steinberg, 1986). Imposing these triangular contrasts on the NR model is the logical equivalent of making the a priori order assumption necessary for the PC model (Thissen, 1988; Masters & Wilson, 1988).

A total of 9652 examinees with usable response strings were obtained from the four administrations; 1985 : N= 2264, 1986 : N= 3026, 1987 : N= 3002, and 1988 : N= 1360. Because of practical and financial considerations the entire data set could not be used and a random sample of 500 examinees with no non-responses was obtained from each administration. Therefore, item parameters for the 10 items were obtained on the responses of 2000 examinees. Because MULTILOG implements a marginal maximum likelihood estimation algorithm the item and person parameters are estimated separately. Therefore, the small number of items relative to the calibration sample's size was not problematic.

The five annual examinations were placed on the same scale by linking the separate exams through the common (1984) items and performing a simultaneous calibration of the four administrations. The crossing of IRT model (PC vs. GR) by rating type (i.e., sum vs. average) produced a 2 X 2 design with one calibration per cell.

Analysis : Analysis consisted of an examination of operating characteristic curves (OCC), item and test information functions. For each rating type the relative efficiency of each of the 10 items when calibrated using the PC model

was compared to those of the items when calibrated using the GR model; the same approach was used for comparing the annual administrations. For each model the relative efficiencies using the sum rating was contrasted with the average rating. Further, the relative efficiencies of the expository items with respect to the narrative items were examined for each model.

Results and Conclusion

Both the PC and GR models could not satisfactorily fit the 1988 administration's narrative item. Therefore, for the following presentation this item as well as the expository item for this testing were eliminated.

Despite the fact that the exams were not developed utilizing a target information function, the inspection of test informations across administrations revealed that, although the functions were not identical, they were relatively similar for the PC model regardless of rating scale used (Figures 1 and 2). Some administrations provided more information in particular θ ranges than others, but, in general, the exams appeared to be measuring ability with relatively the same degree of accuracy. In contrast, for the GR model one can see from Figures 3 and 4 that while the information functions for the 1984-1986 administrations were very similar, the 1987 testing yielded greater information than the other administrations, regardless of rating scale used. In addition, unlike the other administrations the 1986 testing yielded greater information in the approximate θ range of -1.0 to 1.5 based on the average rating scale than it did using the sum rating scale.

Insert Figures 1 to 4 about here

The relative efficiencies of the two rating methods for each model are presented in Figure 5. As can be seen from this figure, for both models the sum rating type provided greater information than the average rating scale. The additional categories in the sum rating scale (relative to the average rating scale) provide greater information than the more restricted average rating scale. In general, the θ range encompassed by the difficulty parameters based on the sum rating scale was larger than for the average

rating scale. For example, for the GR model the average rating scale's difficulty parameter, b_3, had estimates which (roughly) corresponded in magnitude to those of b_5 using the sum rating scale (6 difficulty parameters). Similarly, for the PC model based on the average rating scale the step difficulty estimates for a score of 3 were between the sum rating scale's step difficulties of b_5 and b_6.

Insert Figure 5 about here

As can also be seen from Figure 5, the difference between the PC model's information functions based on the sum and average rating scales was substantial. It was expected that the sum rating scale would provide more information than the average rating scale given that, for the PC model, four-step items have been found to yield more total information across the θ continuum than three-step items (Dodd & Koch, 1987). However, for the GR model the difference between the test information functions based on the two scales was not as dramatic. This lack of a substantial difference between the information functions for the rating scales may be due to the GR model's use of a discrimination parameter. That is, to a certain extent large $\underline{a}$s for items scored using the average scale may compensate for the sum rating scale's additional categories. For the average rating scale the mean $\underline{a}$ was 2.285, whereas for the sum rating scale the average $\underline{a}$ was 2.161. Although two-thirds of the average rating scale items had $\underline{a}$s which were larger than the corresponding sum rating scale items' $\underline{a}$s, this increase in the average rating scale's mean discrimination was primarily the result of three items where the differences were 0.30, 0.37, and 0.51. In those cases where the average rating scale items' $\underline{a}$s were less than those of the sum rating scale, the differences were 0.19, 0.04, 0.04.

In addition, Figure 5 shows that the GR model using the sum ratings provided more information than did the PC model, regardless of rating scale used with the PC model. Differences between models and rating scales for $\theta \geq 3.0$ or

Previous (classical test theory) research with scoring scales has indicated that larger scale ranges produce higher reliabilities than smaller scale ranges (e.g., Coffman, 1971; Godshalk, Swineford, & Coffman, 1966). Results from this study indicated that for either model the larger rating scale (sum holistic rating) yielded greater information than did the smaller rating scale (average holistic rating).

For the PC model using the sum rating scale there were three reversals (out of sequence of bs) for all items. Specifically, an ordering of the step difficulties found the following relationships : $b_2 < b_1 < b_4 < b_3 < b_6 < b_5$. One possible interpretation of this finding is that given a rating of, e.g., a 3 (e.g., based on the b_2/b_1 reversal), it was "easier" or more likely for the examinee to be given a rating of, e.g., a 4; the same logic may be applied to the other reversals. Consistent with this interpretation the PC model's OCC for the 1985 administration's narrative item (Figure 10) showed that certain rating scores were not as probable as others. As can be seen from this figure, the probability of obtaining a rating score of 3 (category score of 1) was far less than that of raw scores of 2 or 4 (category scores of 0 or 2, respectively). Similarly, rating scores of 5 or 7 were not as likely to be given as scores of 4 and 6 or 6 and 8, respectively.

Insert Figure 10 about here

With respect to the GR model and as would be expected, the difficulty parameters for the sum rating scale fitted by the GR model did not exhibit any reversals. That is, the GR model's difficulty parameters correspond to the points of inflection for the category characteristic curves; a set of ogival shape curves specifying the cumulative probabilities of responding in one set of one or more categories versus another set of one or more categories. As stated above, in order to obtain the probability of responding in a particular category the cumulative probabilities for adjacent categories must be subtracted. This fact implies that the difficulty parameters associated with an item's options must be ordered.

Although the GR model does not allow reversals to occur in its difficulty parameters, the model's OCCs for the 1985 administration's narrative item showed a striking similarity to the PC model's for the same item (Figure 11); a comparison of Figures 10 and 11 showed subtle differences between corresponding category scores (e.g., 0, 1, 5, and 6). In fact, if the GR model's difficulty parameters were defined as those of the PC model, then reversals would have occurred for the GR model as well.

Insert Figure 11 about here

Because the interpretation of the GR model's OCCs parallels that of the PC model's it appears that, in effect, for both models the 7-point rating scale became functionally a 4 point scale.

This final analysis demonstrated how either model may be used for a rating scale analysis. Unlike classical techniques, this method is based on parameter estimates which are not sample dependent. In addition, although this study's data collection method precluded an analysis of interrater agreement, it is possible, if data collection is structured appropriately, to perform an interrater agreement analysis through the use of item or test information functions; if desired, items may be grouped to form "concept" information functions and the interrater analysis may be performed on a concept basis. In this regard, one could not only determine the degree of interrater agreement, but also where there was a lack of agreement. It is felt that the advantages of IRT methods may be realized with essay-type examinations.

$\theta \leq -3.0$ are not considered meaningful given the minimal amount of information provided by each model in these ability regions. In general, the interaction of relatively large discrimination parameters (with respect to the assumed constant a value in the PC model) and a wider θ range encompassed by the GR difficulty parameters than by the PC model's, resulted in the GR model providing greater information than did the PC model. The GR model's test information function based on the sum rating provided more information than the PC model's test information function with the same rating scale.

Figure 6 shows that for the majority of the θ continuum the expository items provided greater information than the narrative items when they were fitted by the PC model. An item analysis by administration showed that, except for the 1985 administration ($-0.75 \leq \theta \leq 2.0$), as θ increased the expository items provided more information than did the narrative items. These relationships are presented in Figure 7. As can be seen, for higher ability examinees expository items provided increases in information of as much as $1^1/_2$ times that of narrative items.

Insert Figures 6 and 7 about here

In contrast to the PC model's results, expository items fitted by the GR model only provided information greater than the narrative items for θs greater than approximately 1.0 (Figure 6). An inspection of the relative efficiencies for the narrative versus expository items per administration (Figure 8) revealed that, except for (as was the case for the PC model) the 1985 testing, the pattern of expository items' information increasing with increasing θ is also evident for the GR model. However, under the GR model the narrative items appear to provide more information than the expository items for a larger portion of θ continuum than they did under the PC model. For both models, results based on the average rating scale were similar to those of the sum rating scale.

Insert Figure 8 about here

Discussion

The results of this study appear to indicate that both models are useful for the calibration of writing samples. However, for this item set the GR model provided more information than the PC model for both rating scales. The greater information provided by the GR model was primarily a result of a discrimination parameter which was allowed to vary across items ($1.789 \leq \underline{a} \leq 2.823$) and which in all cases was larger than the assumed constant $\underline{a}$ value of the PC model. In those cases where $\underline{a}$ is relatively constant across items (e.g., fitting a GR model and examining the item discriminations), one may prefer to use the PC model because of the fewer parameters to estimate and the minimal gains to be expected by using the GR model in this context. In addition, the decision to use one model over the other may be a result of pragmatic constraints as well as philosophical beliefs concerning the estimation of parameters other than $\underline{b}$ (see Wright and Stone (1979) for more information on the estimation controversy).

Results indicated that, in general, expository items provided more information for higher θs than narrative items did. This finding is consistent with the nature of these two discourse models. Therefore, for the assessment of high abilities, items of an expository nature would be preferred to narrative items. Of course, after a sufficiently large item pool is developed and calibrated, item information functions may be utilized to construct exams according to a target test information function and/or which are essentially weakly parallel (Samejima, 1977). For instance, Figure 9 contains the information functions for the most informative test actually administered (i.e., the 1987 administration) and a hypothetical four-item test (constructed from the 8-item pool used in this study). As can be seen the developed test (items : 1985/1987 expository and 1985/1986 narrative items) provided almost twice the information as the 1987 administration in addition to providing more of this information over a wider θ range.

Insert Figure 9 about here

References

Ackerman, T. (1986). Use of the graded response IRT model to assess the reliability of direct and indirect measures of writing assessment. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April, 1986.

Bock, R.D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37, 29-51.

Breland, H.M. (1983). The direct assessment of writing skill : A measurement review. ETS Research Report (86-9). Princeton, NJ : Educational Testing Service.

Coffman, W.E. (1971). Essay examinations. In Educational Measurement. R.L. Thorndike (Ed). Washington, D.C. : American Council on Education.

Dodd, B.G. & Koch, W.R. (1987). Effects of variations in item step values on item and test information in the partial credit model. Applied Psychological Measurement, 4, 371-384.

Godshalk, F.I. Swineford, F., & Coffman, W.E. (1966). The Measurement of Writing Ability. New York: College Entrance Examination Board.

Masters, G. (1982). A Rasch model for partial credit scoring. Psychometrika, 47, 149-174.

Masters, G. & Wilson, M. (1988). Understanding and using partial credit analysis : an IRT method for ordered response categories. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, April 16-20.

Pollitt, A. & Hutchinson, C. (1987). Calibrating graded assessments: Rasch partial credit analysis of performance in writing. Language Testing, 4, 72-92.

Samejima, F. (1977). A use of the information function in tailored testing. Applied Psychological Measurement, 1, 233-247.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph Supplement, (No. 17).

Stiggins, R.J. & Bridgeford, N.J. (1983). An analysis of published tests of writing proficiency. Educational Measurement : Issues and Practice, 2, 6-10 & 26.

Thissen, D.J. (1988). MULTILOG-User's Guide (Version 5.1). Scientific Software, Inc. Mooresville, IN.

Thissen, D.J. & Steinberg, L. (1986). A taxonomy of item response models. Psychometrika, 51, 567-577.

Wright, B.D. & Stone, M.H. (1979). Best Test Design. Chicago: MESA Press.

Figure 1

Test Information Functions for 1984-1988 administrations for PC model
Sum Rating Scale

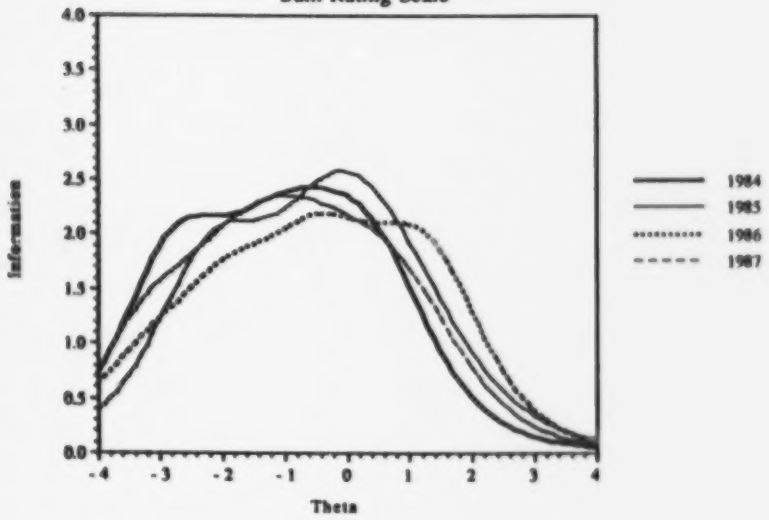

Figure 2

Test Informations for 1984-1988 administrations for PC model
Average Rating Scale

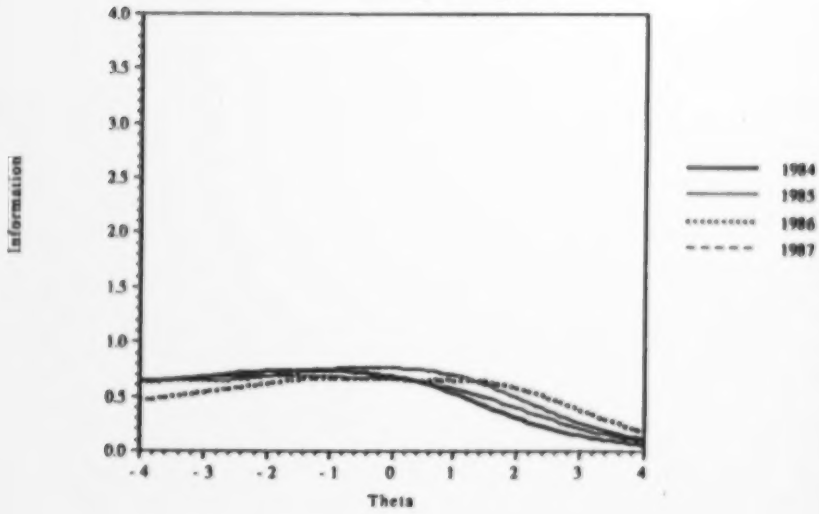

Figure 3

Test Informations for 1984-1987 administrations for GR model
Sum Rating Scale

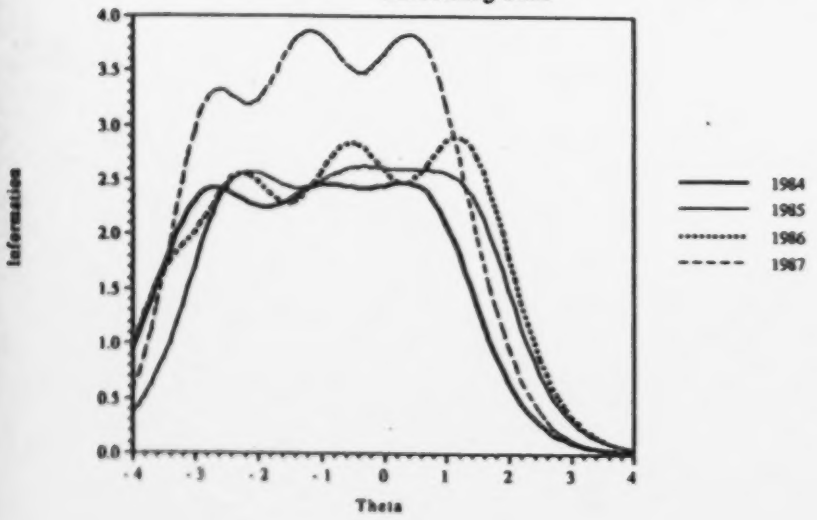

Figure 4

Test Informations for 1984-1987 administrations for GR Model
Average Rating Scale

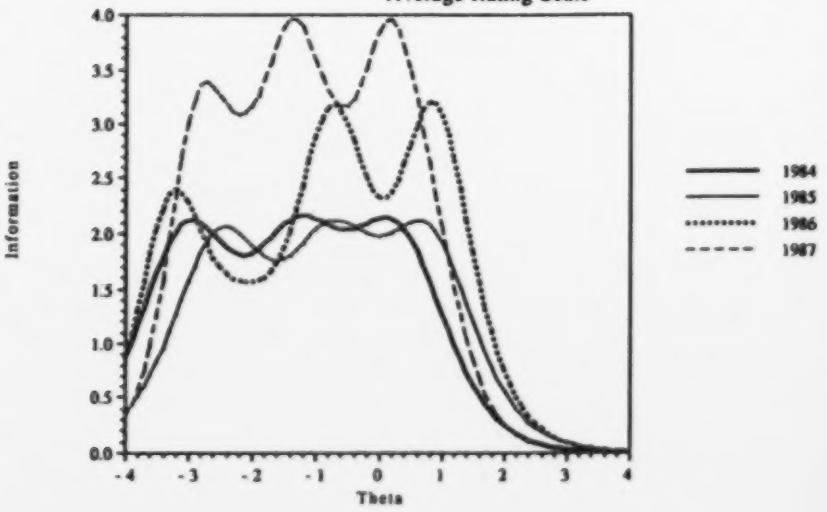

Figure 5

Relative Efficiency of Rating Scales for the PC & GR models

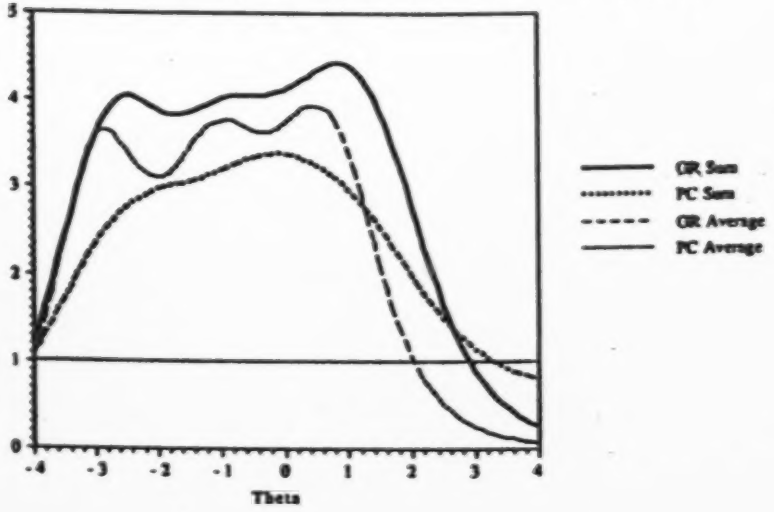

Figure 6

Information Functions for Narrative & Expository Item Types for Sum Rating

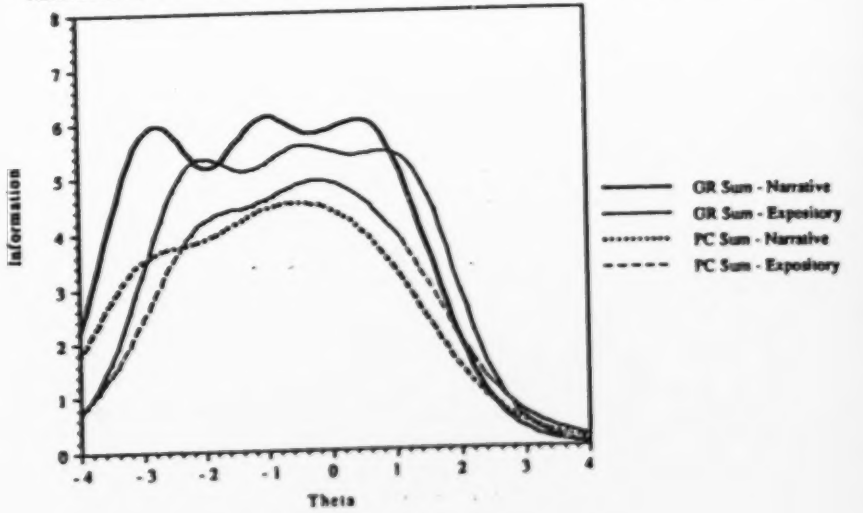

Figure 7

Relative Efficiencies for Expository vs. Narrative Items
PC Model Sum Rating Scale

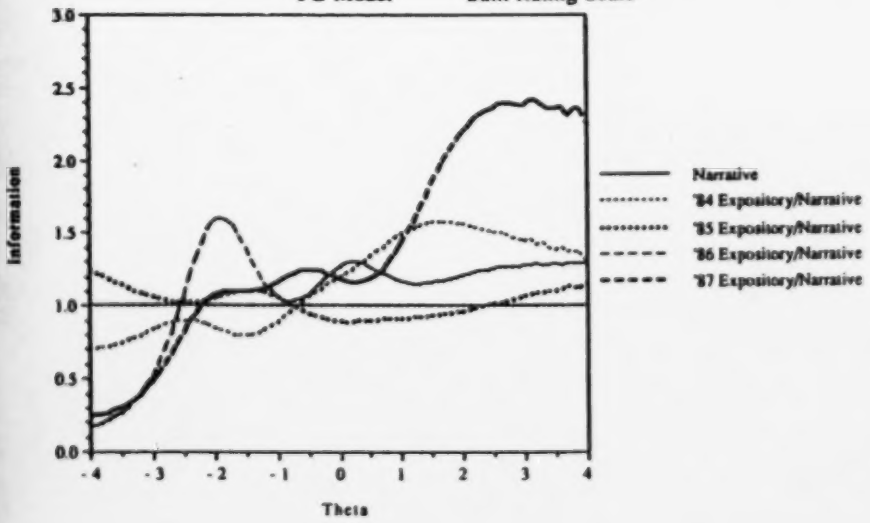

Figure 8

Relative Efficiencies for Expository vs. Narrative Items
GR Model Sum Rating Scale

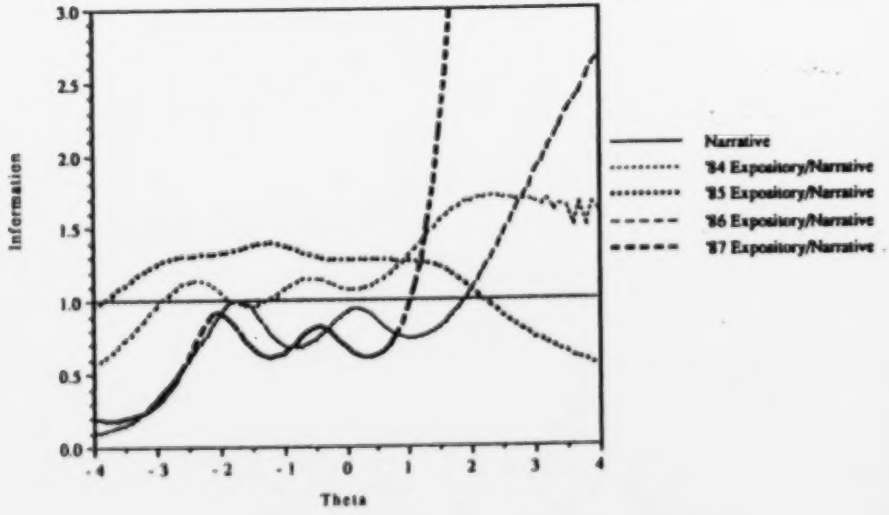

Figure 9

Test Development

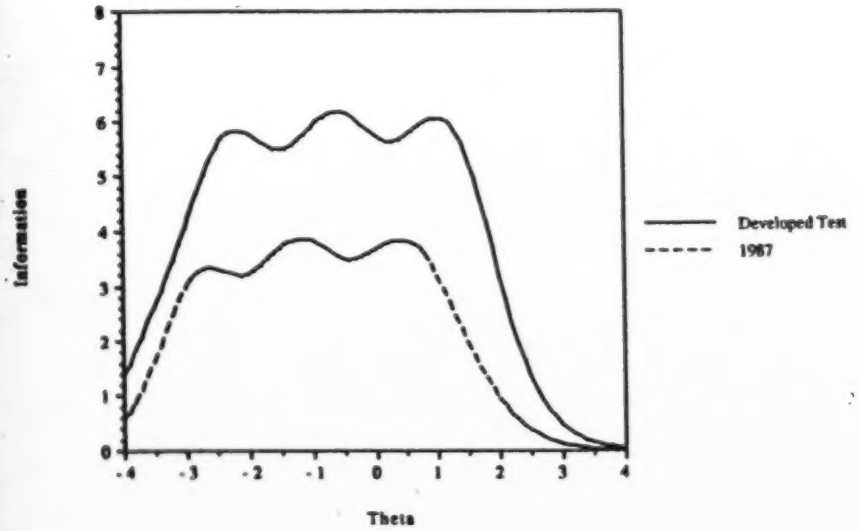

Figure 10

1985 Narrative Item - PC Model

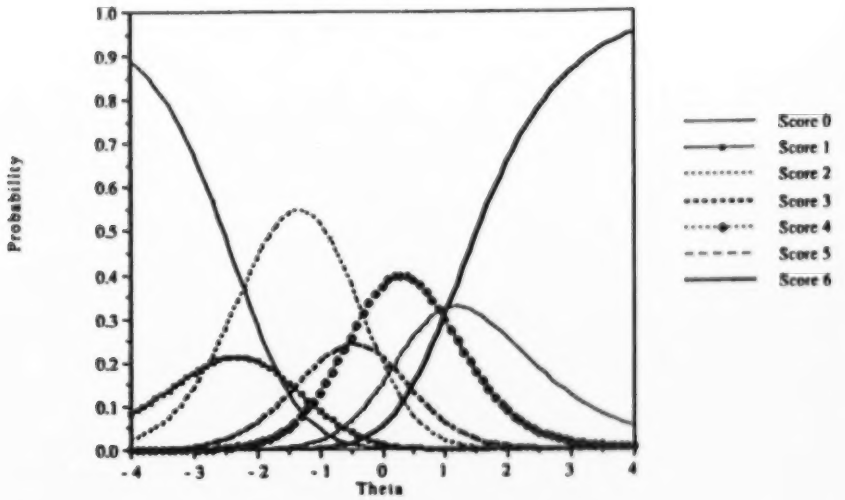

Figure 11

1985 Narrative Item - GR Model